Shapley Curves: A New Concept for Modelling Feature Importance

Farjad Adnan, Karlson Pfannschmidt, Eyke Hüllermeier

Intelligent Systems Group
Paderborn University

We propose a novel method for measuring the importance and usefulness of predictor variables (features) in supervised machine learning, which makes use of concepts from cooperative game theory. The basic idea of our approach is to consider subsets of variables as coalitions, and their predictive performance as a payoff. This approach acknowledges the fact that the usefulness of a feature in a learning context strongly depends, not only on the learning method being used, but also on the other features being available.

A theoretically appealing measure of the importance of an individual feature is the *Shapley value* [3]. Computationally, however, this measure is challenging. First, the exact computation of the Shapley values requires determining the performance of all possible subsets of features, which is in general #P-hard [2]. Furthermore, in the context of machine learning, even the training of a single predictor on one subset of features can take a considerable amount of time.

As another aspect specific to machine learning, let us note that the Shapley values of each feature can change with varying sample size, due to effects such as overfitting. Motivated by this observation, we introduce the concept of a *Shapley curve*, which depicts the (weighted average) contribution of a feature to the learning curve (expected performance as a function of the sample size).

We develop an approximation technique for estimating Shapley values, which is efficient in the number of models that need to be trained and validated. Moreover, to estimate Shapley curves, we propose a hierarchical Bayes approach that does not require an evaluation of all possible subsets of features on different sample sizes. Last but not least, leveraging related techniques for extrapolating learning curves [1], we are able to estimate the Shapley values in the limit when the sample size goes to infinity. We evaluate our approach on synthetic and real-world datasets.

References

1. C. Cortes, L.D. Jackel, S.A. Solla, V. Vapnik, and J.S. Denker. Learning curves: Asymptotic values and rate of convergence. In *Proc. NIPS, Advances in Neural Information Processing Systems*, Denver, USA, 1993.
2. X. Deng and C.H. Papadimitriou. On the complexity of cooperative solution concepts. *Math. Oper. Res.*, 19(2):257–266, 1994.
3. K. Pfannschmidt, E. Hüllermeier, S. Held, and R. Neiger. Evaluating tests in medical diagnosis: Combining machine learning with game-theoretical concepts. In *Proc. IPMU, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Eindhoven, The Netherlands, 2016.

In: M. Leyer (Ed.): Proceedings of the LWDA 2017 Workshops: KDML, FGWM, IR, and FGDB.
Rostock, Germany, 11.-13. September 2017, published at http://ceur-ws.org